

งานคำศัพท์

1. keywords

a. Bigdata

- i. ชุดข้อมูลที่มีขนาดใหญ่ ความหลากหลายสูง และความเร็วในการเกิดขึ้นสูงมากจนการจัดการด้วยเทคโนโลยีแบบดั้งเดิมไม่สามารถทำได้อย่างมีประสิทธิภาพ

b. Data Warehouse

- i. เป็นระบบที่ใช้เก็บรวบรวมข้อมูลจากแหล่งต่างๆ ภายในองค์กรเพื่อใช้ในการวิเคราะห์และทำรายงาน

c. Data Lake

- i. เป็นที่เก็บข้อมูลขนาดใหญ่ที่สามารถเก็บข้อมูลทุกประเภท

d. Lakehouse

- i. เป็นสถาปัตยกรรมที่รวมคุณสมบัติของ **Data Warehouse** และ **Data Lake** เข้าไว้ด้วยกัน โดยมอบความยืดหยุ่นในการจัดการข้อมูลทุกประเภท

e. ACID Transaction

- i. **Atomicity, Consistency, Isolation, Durability** เป็นคุณสมบัติที่สำคัญของระบบจัดการฐานข้อมูลซึ่งช่วยให้มั่นใจได้ว่าการดำเนินการบนข้อมูลนั้นถูกต้องและเชื่อถือได้

1. **Atomicity:** ทุกการดำเนินการ (Transaction) จะต้องสำเร็จทั้งหมดหรือไม่สำเร็จเลย

2. **Consistency:** การดำเนินการจะต้องทำให้ฐานข้อมูลอยู่ในสถานะที่ถูกต้องเสมอ

3. **Isolation:** การดำเนินการต่างๆ จะต้องไม่ส่งผลกระทบต่อกัน

4. **Durability:** ข้อมูลที่บันทึกไปแล้วจะต้องไม่หายไปแม้ว่าจะเกิดความล้มเหลวในระบบ

f. Data Scientist

- i. ผู้เชี่ยวชาญในการวิเคราะห์และตีความข้อมูลจำนวนมาก เพื่อช่วยให้องค์กรสามารถตัดสินใจได้อย่างมีข้อมูลรองรับ

g. Data Engineer

- i. ผู้ที่ออกแบบ, สร้าง, และดูแลโครงสร้างพื้นฐานของข้อมูล (Data Infrastructure) รวมถึงการสร้างกระบวนการสำหรับการเก็บ, การประมวลผล, และการส่งข้อมูลให้กับ Data Scientist หรือระบบอื่นๆ

h. Apache Spark

- i. ระบบประมวลผลข้อมูลแบบกระจาย (Distributed Data Processing Framework) ที่ใช้ในการประมวลผลข้อมูลขนาดใหญ่ได้อย่างรวดเร็ว

i. Spark SQL

- i. โมดูลของ Apache Spark ที่ใช้สำหรับการประมวลผลข้อมูลในรูปแบบโครงสร้าง (Structured Data) Spark SQL มอบความสามารถในการเขียนโค้ด SQL เพื่อทำการดึงข้อมูล, การกรอง, และการวิเคราะห์ข้อมูล

2. Big Data

a. Volume

- i. ปริมาณข้อมูลที่มีขนาดใหญ่มหาศาล ข้อมูลเหล่านี้สามารถมาจากหลายแหล่ง

b. Velocity

- i. ความเร็วในการเกิดขึ้นและการประมวลผลของข้อมูล Big Data มีความเร็วในการเกิดขึ้นสูง

c. Variety

- i. ความหลากหลายของรูปแบบข้อมูล ข้อมูลใน Big Data มักจะมีทั้งข้อมูลที่มีโครงสร้าง (Structured), ไม่มีโครงสร้าง (Unstructured), และกึ่งโครงสร้าง (Semi-Structured)

d. Veracity

- i. ความถูกต้องและความน่าเชื่อถือของข้อมูล ใน Big Data ข้อมูลอาจมีความไม่แน่นอนหรือความไม่ถูกต้องสูง

e. Value

- i. คุณค่าหรือประโยชน์ที่สามารถสกัดออกมาจากข้อมูล Big Data

3. Data Warehouse / Data Lake

a. Data Warehouse

i. Structured

1. ข้อมูลใน Data Warehouse มักจะมีโครงสร้างชัดเจนและเป็นข้อมูลที่ผ่านการปรับปรุงมาแล้ว

ii. การจัดระเบียบข้อมูล

1. ข้อมูลใน Data Warehouse จะถูกจัดเก็บในรูปแบบที่สามารถใช้ในการวิเคราะห์และการทำงานได้อย่างมีประสิทธิภาพ

iii. การประมวลผลข้อมูล

1. มักใช้สำหรับการประมวลผลข้อมูลในเชิงวิเคราะห์

iv. Amazon Redshift, Snowflake Elastic Data Warehouse

b. Data Lake

i. Unstructured

1. **Data Lake** สามารถเก็บข้อมูลทุกประเภท ทั้งข้อมูลที่มีโครงสร้าง, ไม่มีโครงสร้าง, และกึ่งโครงสร้าง

ii. ความยืดหยุ่นสูง

1. ข้อมูลสามารถถูกเก็บในรูปแบบดั้งเดิมโดยไม่ต้องปรับปรุง

iii. การประมวลผลข้อมูล

1. **Data Lake** ถูกใช้ในการประมวลผลข้อมูลแบบเรียลไทม์หรือการวิเคราะห์ข้อมูลเชิงลึก และการเรียนรู้ของเครื่อง

iv. **Azure Data Lake Storage, Amazon S3**

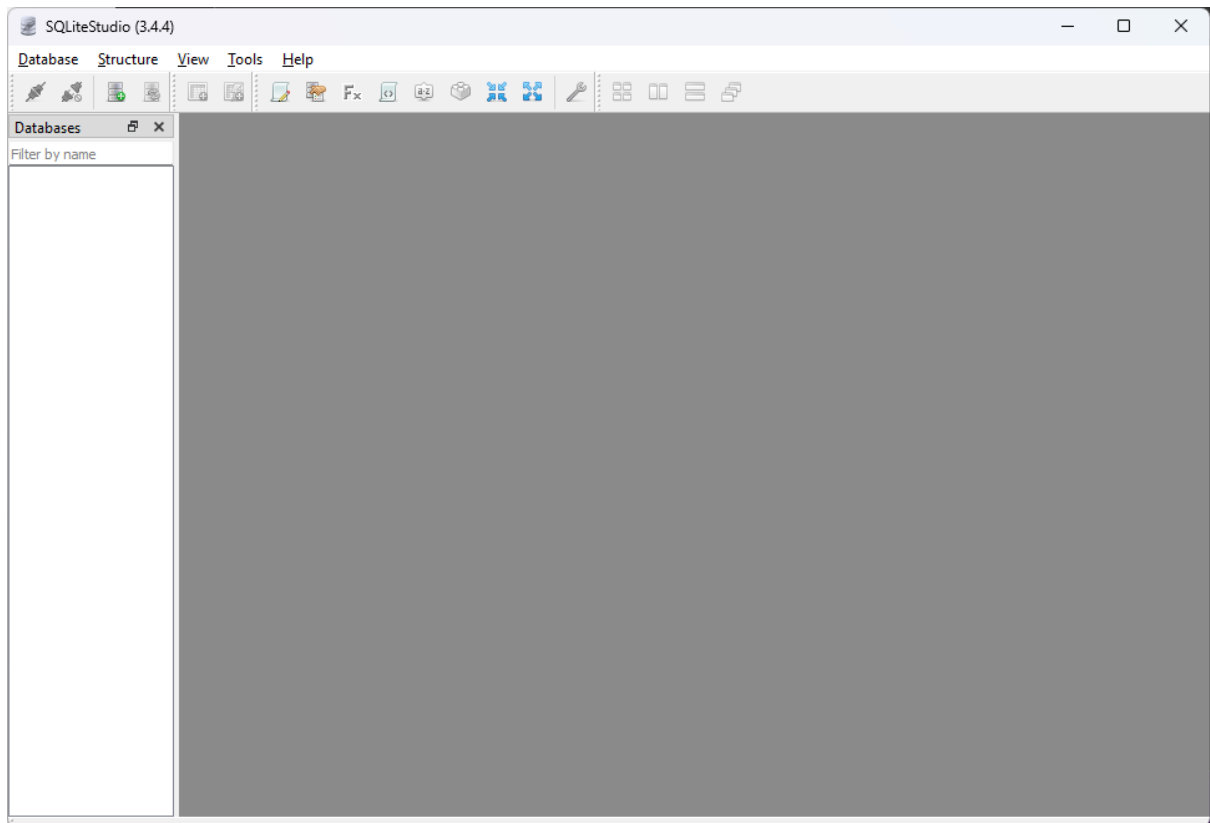
SQL for Data Science

1. Download & Install

```
Command Prompt - sqlite3
C:\sqlite>sqlite3
SQLite version 3.46.0 2024-05-23 13:25:27 (UTF-16 console I/O)
Enter ".help" for usage hints.
Connected to a transient in-memory database.
Use ".open FILENAME" to reopen on a persistent database.
sqlite> |
```

```
Command Prompt
.parameter CMD ...      Manage SQL parameter bindings
.print STRING...        Print literal STRING
.progress N             Invoke progress handler after every N opcodes
.prompt MAIN CONTINUE   Replace the standard prompts
.quit                  Stop interpreting input stream, exit if primary.
.read FILE              Read input from FILE or command output
.recover                Recover as much data as possible from corrupt db.
.restore ?DB? FILE      Restore content of DB (default "main") from FILE
.save ?OPTIONS? FILE    Write database to FILE (an alias for .backup ...)
.scanstats on|off|est   Turn sqlite3_stmt_scanstatus() metrics on or off
.schema ?PATTERN?       Show the CREATE statements matching PATTERN
.separator COL ?ROW?    Change the column and row separators
.session ?NAME? CMD ... Create or control sessions
.sha3sum ...            Compute a SHA3 hash of database content
.shell CMD ARGS...      Run CMD ARGS... in a system shell
.show                  Show the current values for various settings
.stats ?ARG?            Show stats or turn stats on or off
.system CMD ARGS...     Run CMD ARGS... in a system shell
.tables ?TABLE?         List names of tables matching LIKE pattern TABLE
.timeout MS             Try opening locked tables for MS milliseconds
.timer on|off           Turn SQL timer on or off
.trace ?OPTIONS?        Output each SQL statement as it is run
.version                Show source, library and compiler versions
.vfsinfo ?AUX?          Information about the top-level VFS
.vfslist                List all available VFSes
.vfsname ?AUX?          Print the name of the VFS stack
.width NUM1 NUM2 ...    Set minimum column widths for columnar output
sqlite> .quit

C:\sqlite>|
```



2. Connect to SQLite sample database

```
Command Prompt - sqlite3 c
C:\sqlite>sqlite3 c:\sqlite\db\chinook.db
SQLite version 3.46.0 2024-05-23 13:25:27 (UTF-16 console I/O)
Enter ".help" for usage hints.
sqlite> .table
albums          employees       invoices        playlists
artists         genres         media_types    tracks
customers       invoice_items  playlist_track
sqlite> |
```

SQLiteStudio (3.4.4) - [SQL editor 1]

Database Structure View Tools Help

Databases Filter by name

chinook (SQLite 3)

- Tables (11)
 - albums
 - artists
 - customers
 - employees
 - genres
 - invoice_items
 - invoices
 - media_types
 - playlist_track
 - playlists
 - tracks
- Views

Query History

1 select * from albums;

Grid view Form view

Total rows loaded: 347

	AlbumId	Title	ArtistId
1	1	For Those About To Rock We Salute You	1
2	2	Balls to the Wall	2
3	3	Restless and Wild	2
4	4	Let There Be Rock	1
5	5	Big Ones	3
6	6	Jagged Little Pill	4
7	7	Facelift	5
8	8	Warner 25 Anos	6
9	9	Plays Metallica By Four Cellos	7
10	10	Audioslave	8

Status [17:02:59] Query finished in 0.000 second(s).

albums (chinook) artists (chinook) SQL editor 1